

SELF-ALIGNED SPLIT-GATE NAND FLASH MEMORY AND FABRICATION PROCESS

Background of the Invention

Field of Invention

This invention pertains generally to semiconductor memory devices and, more particularly, to a self-aligned split-gate NAND flash memory and process of
5 fabricating the same.

Related Art

Nonvolatile memory is currently available in several forms, including electrically
programmable read only memory (EPROM), electrically erasable programmable
10 read only memory (EEPROM), and flash EEPROM. Flash memory has been
widely used for high volume data storage in devices such as memory cards,
personal digital assistants (PDA's), cellular phones, and MP3 players. Such
applications require high density memory, with smaller cell size and reduced cost
of manufacture.

15 NOR-type stack-gate flash memory cells typically have a bit line contact, a
source region, a floating gate, and a control gate, with the control gate being
positioned directly above the floating gate. The relatively size of such cells
prevents them from being used in very high density data storage applications.

20 Cell size is smaller in a NAND flash memory array having a series of stack-gate
flash memory cells connected in series between a bit-line and a source line, with
only one bit-line contact. Such an array is illustrated in Figure 1 and described in
greater detail in U.S. Patents 4,959,812 and 5,050,125. In this array, stack-gate
25 memory cells 21 are connected in series between a bit line 22 and a source line
23. The cells are formed in a P-well 24 in a substrate 26 of either N- or P-type
silicon. Each of the cells has a floating gate 27 fabricated of a conductive
material such as polysilicon and a control gate 28 fabricated of a conductive

material such as polysilicon or polycide. The control gate is positioned above and in vertical alignment with the floating gate.

Two select gates 29, 31 are included in the array, one near the bit line contact 32 and one near source diffusion 23. Diffusions 33 are formed in the substrate between the stacked gates and between the stacked gates and the select gates to serve as source and drain regions for the transistors in the memory cells. Bit line diffusion 22, source diffusion 23, and diffusions 33 are doped with N-type dopants.

To erase the memory cell, a positive voltage of about 20 volts is applied between the P-well and the control gates, which causes the electrons to tunnel from the floating gates to the channel regions beneath them. The floating gates thus become positively charged, and the threshold voltage of the stack-gate cells becomes negative.

To program the memory cells, the control gates are biased to a level of about 20 volts positive relative to the P-well. As electrons tunnel from the channel region to the floating gates, the floating gates are negatively charged, and the threshold voltage of the stack-gate cells becomes positive. By changing the threshold voltage of a stack-gate cell, the channel beneath it can be in either a non-conduction state (logical "0") or a conduction state (logical "1") when a zero voltage is applied to the control gate during a read operation.

However, as fabrication processes advance toward smaller geometries, *e.g.* tens of nanometers, it is difficult to form a high-voltage coupling ratio which is sufficient for program and erase operations while maintaining a small cell size and meeting stringent reliability requirements such as 10-year data retention and 1,000,000 cycling operations between failures.

Objects and Summary of the Invention

It is in general an object of the invention to provide a new and improved semiconductor device and process for fabricating the same.

Another object of the invention is to provide a semiconductor device and process of the above character which overcome the limitations and disadvantages of the prior art.

These and other objects are achieved in accordance with the invention by providing a self-aligned split-gate NAND flash memory cell array and process of fabrication in which rows of self-aligned split-gate cells are formed between a bit line diffusion and a common source diffusion in the active area of a substrate.

5 Each cell has control and floating gates which are stacked and self-aligned with each other, and erase and select gates which are split from and self-aligned with the stacked gates, with select gates at both ends of each row which partially overlap the bit line the source diffusions. The channel regions beneath the erase

10 gates are heavily doped to reduce the resistance of the channel between the bit line and source diffusions, and the floating gates are surrounded by the other gates in a manner which provides significantly enhanced high voltage coupling to the floating gates from the other gates. The array is biased so that all of the memory cells in it can be erased simultaneously, while programming is bit

15 selectable.

Brief Description of the Drawings

Figure 1 is a cross-sectional view of a NAND flash memory array with a series of stack-gate flash memory cells of the prior art.

20

Figure 2 is a cross-sectional view, taken along line 2 – 2 in Figure 3, of one embodiment of a self-aligned split-gate NAND flash memory cell array incorporating the invention.

25 Figure 3 is a top plan view of the embodiment of Figure 2.

Figures 4A – 4H are schematic cross-sectional views illustrating the steps in the embodiment of a process for fabricating a NAND flash memory cell array in accordance with the invention.

30

Figures 5A – 5D are circuit diagrams of a small memory array as in one embodiment of Figure 2, showing exemplary bias conditions for erase, program and read operations.

35 Figure 6 is a cross-sectional view, taken along line 6 – 6 in Figure 7, of another embodiment of a self-aligned split-gate NAND flash memory cell array incorporating the invention.

Figure 7 is a top plan view of the embodiment of Figure 6.

5 Figures 8A – 8E are schematic cross-sectional views illustrating the steps in one embodiment of a process for fabricating a NAND flash memory cell array in accordance with the invention.

10 Figures 9A – 9B are circuit diagrams of a small memory array as in the embodiment of Figure 6, showing exemplary bias conditions for erase, program and read operations.

Detailed Description

15 As illustrated in Figures 2 and 3, the memory includes an array of split-gate NAND flash memory cells 36, each of which has a floating gate 37 and a control gate 38, with the control gate being positioned above and in vertical alignment with the floating gate.

20 The floating gates are relatively thin and are fabricated of a conductive material such as polysilicon or amorphous silicon, with a preferred thickness on the order of 100Å to 1000Å. Thin gate insulators 40, typically a thermal oxide, are positioned between the floating gates and the underlying the substrate.

25 The control gates are narrower in horizontal dimension and thicker in vertical dimension than the floating gates, with the edge portions of the floating gates extending laterally beyond the edge portions of the control gates. The control gates are fabricated of a conductive material such as a doped polysilicon or polycide, and each of the control gates is insulated from the floating gate beneath it by a dielectric film 42. That film can be either a pure oxide or a combination of oxide, nitride and oxide (ONO), and in the presently preferred
30 embodiment, it consists of a layer nitride between two layers of oxide.

35 Erase gates 43 and select gates 44 are disposed alternately between stack-gate cells 36, and additional select gates 44a, 44b are adjacent to the cells at the ends of the group. These gates are fabricated of a conductive material such as a doped polysilicon or polycide, and are self-aligned with and parallel to the adjacent control gates and floating gates, with thick dielectric films 47 separating them from the adjacent control gates and thin tunnel oxides 48 separating them

from the floating gates. Both the dielectric films and the tunnel oxides can be either a pure thermal oxide or a combination of thermal oxide, a CVD oxide and a CVD nitride.

5 Diffusion regions 49, a bit line diffusion 50, and a common source diffusion 51 are formed in a P-type well 52 in the upper portion of substrate 41 and doped with an N-type material. Diffusion regions 49 are positioned directly beneath the erase gates 43, and bit line diffusion 50 is partially overlapped by the select gate 44a at one end of the array. Common source diffusion region 51 is partially
10 overlapped by the select gate 44b at the other end of the array and is shared by this array of cells and by another array (not shown) of the same type.

Erase gates 43 and select gates 44 are separated from the diffusion regions and the substrate by gate oxides 53, and select gates 44a, 44b are separated from
15 the substrate by gate oxides 54. Oxide layers 53 and 54 can be either a pure thermal oxide or a combination of thermal oxide and CVD oxide.

In this embodiment, erase paths are formed between the side walls 39 of the floating gates and the adjacent erase gates 43 and select gates 44, 44a, 44b,
20 through tunnel oxides 48.

As illustrated in Figure 3, isolation regions 56 are formed between rows of cells, and control gates 38 cross over floating gates 37 and the isolation regions. Erase gates 43 and select gates 44, 44a, 44b are parallel to the control gates.
25 Bit lines 57 are perpendicular to those gates, and cross over the bit line contact, gates in the respective rows, and the common source region.

The N+ diffusions 49 beneath the erase gates significantly reduce the resistance of the channel region between bit line diffusion 50 and common source diffusion
30 51. As a result, bit line and common source voltages can pass to selected cells with substantially less voltage drop along the channel. This permits the length of the structure and the number of cells in each row to be substantially greater than in devices without the N+ diffusions, *e.g.* 32 cells vs. 16.

35 The memory cell array of Figures 2 and 3 can be fabricated by the process illustrated in Figures 4A through 4H. In this process, an oxide layer 58 is thermally grown to a thickness of about 70Å to 150Å on a monocrystalline silicon

substrate which is illustrated as comprising a P-type substrate 41 in which a P-type well 52 is formed. Alternatively, if desired, an N-type well can be formed in the P-type substrate, in which case the P-type well is formed in the N-type well.

5

A conductive layer 59 of polysilicon or amorphous silicon (poly-1) is deposited on the thermal oxide to a thickness on the order of 100Å to 1000Å, and a dielectric layer 61 (the inter-poly dielectric) is formed on the silicon. This silicon is preferably doped with phosphorus, arsenic or boron to a level on the order of 10^{17} to 10^{20} per cm^3 and is subsequently etched to form floating gates 37. The doping can be done *in-situ* during deposition of the silicon or by ion implantation directly into the silicon or through the dielectric 61 above it.

10

The inter-poly dielectric can be either a pure oxide or a combination of oxide, nitride and oxide (ONO), and in the embodiment illustrated, it consists of a lower oxide layer having a thickness on the order of 30 – 100Å, a central nitride layer having a thickness on the order of 60 – 300Å, and an upper oxide layer having a thickness on the order of 30 – 100Å.

15

A second layer of polysilicon 62 (poly-2) is deposited on dielectric film 61, and is subsequently etched to form the control gates 38. This layer has a thickness on the order of 1500Å – 3500Å, and is doped with phosphorus, arsenic or boron to a level on the order of 10^{20} to 10^{21} per cm^3 .

20

A CVD oxide or nitride layer 63 having a thickness on the order of 300Å - 1000Å is deposited on the poly-2 layer, and is used as a mask to prevent the poly-2 material from etching away during subsequent dry etching steps.

25

A photolithographic mask 65 is formed over layer 63 to define the control gates, and the unmasked portions of that layer and the poly-2 layer are etched away anisotropically, leaving only the portions of the poly-2 which form the control gates 38, as illustrated in Figure 4B.

30

The photoresist is then stripped away, and oxide 47 is thermally grown on the side walls of the control gates to a thickness on the order of 200 Å – 700Å, as shown in Figure 4C.

35

Using oxide 47 as a mask, the exposed portions of the inter-poly dielectric 61 and the underlying portions of the poly-1 layer 59 are etched away anisotropically to form floating gates 37, with only a thin layer of oxide 58 being left on the surface of the substrate between the gates.

5

A photolithographic mask 66 is formed to define diffusion regions 49 between every other pair of stack-gate memory cells 36, as illustrated in Figure 4D. Diffusion regions 49 are then formed in the substrate between those gates by ion implantation, using dopants such as P³¹ or As⁷⁵.

10

Following ion implantation, the photoresist is stripped away, and another thermal oxidation is performed, which builds up tunnel oxide 48, thermal oxide 53, and gate oxide 54, as shown in Figure 4E. Tunnel oxide 48 is thus built up to a thickness on the order of 100Å – 250Å, and Gate oxide 54 is built up to a

15

To improve the quality of the oxide films and reduce disturbances between the floating gates and the select and erase gates, a thin CVD oxide of about 50Å - 200Å can be deposited before or after thermal oxidation.

20

As a result of these processing steps, each of the control gates is self-aligned to the floating gate beneath it, the control gate is narrower than the floating gate, and the edge portions of the floating gate extend laterally beyond the edge portions of the control gate.

25

Following thermal oxidation, a conductive layer (poly-3) 64 is deposited over the entire wafer, as illustrated in Figure 4E. This layer is typically doped polysilicon or polycide, and it is deposited to a thickness on the order of 1500Å – 4000Å.

30

The poly-3 layer is then etched anisotropically, leaving only the portions which form erase gates 43 and select gates 44a, 44b, 44, as illustrated in Figure 4F. Being formed in this manner, the erase gates and the select gates are self-aligned with and parallel to the control gates.

35

N-type dopants such as P³¹ or As⁷⁵ are implanted into P-well 52 to form the bit line diffusion 50 and common source diffusion 51, as illustrated in Figure 4G,

with the portions 72 of the P-well 52 beneath select gates 44 being used as the channels for those gates.

5 Thereafter, a glass material such as phosphosilicate glass (PSG) or borophosphosilicate glass (BPSG) is deposited across the entire wafer, then etched to form wells for bit line contacts 46. Finally, a metal layer is deposited over the glass and patterned to form bit lines 57 and bit line contacts 46.

10 Operation and use of the memory cell array can be described with reference to Figures 5A – 5D where exemplary bias voltages for erase (ERS), program (PGM) and read (RD) operations are shown next to the terminals of the array. In the examples of Figures 5A – 5C, the selected memory cell is C_{1n} which is located at the intersection of bit line BL_n and control gate CG_1 . In the example of Figure 5D, the selected cell is C_{2n} located at the intersection of bit line BL_n and
15 control gate CG_2 . The selected cells are circled for ease of location. All of the other memory cells in the array are unselected during PGM and RD operations.

20 During an erase operation, electrons are forced to tunnel simultaneously from the floating gates to neighboring erase gates and select gates for all cells in array, leaving positive the floating gates positively charged. When the electric field across the tunnel oxide is more than about 10^7 V/cm, Fowler-Nordheim tunneling becomes significant, and electrons with sufficient energy can tunnel from the cathode electrode (floating gate) to the anode electrodes (erase gate and select gate).

25 Erasing can be done using either of two bias conditions. In the first erase mode, the control gates of the memory cells are biased at -7 to -12 volts, the select gates SG_0 to SG_{16} and the erase gates EG_0 – EG_{15} are biased at 3 – 7 volts, and the bit lines and the common source are floating. In the second mode, the
30 control gates are biased at 0 volts, the select gates SG_0 to SG_{16} and erase gates EG_0 – EG_{15} are biased at 9 – 12 volts, the P-well 52 is biased at 0 volt, and the bit lines and the common source are floating.

35 With these bias conditions, most of the voltage difference between the control gates and the select gates or erase gates appears across the tunnel oxides surrounding the side walls of floating gates. That triggers Fowler-Nordheim tunneling, with electrons tunneling from the floating gates to adjacent select

gates and erase gates for all cells in array. As the floating gates become more positively charged, the threshold voltages of the memory cells, which are preferably in the range of -2 to -5 volts, become lower. This results in an inversion layer in the channel under the floating gate when the control gate is biased at 0 volts. Therefore, the memory cell goes into the conductive state (logic "1") after the erase operation. In an unselected array, the control gates and the erase gates are biased at 0 volts, and there is no Fowler-Nordheim tunneling during the erase operation.

10 In the program mode shown in Figure 5A, the control gate CG_1 of the selected memory cell C_{1n} is biased to a level of 15 – 18 volts; 5 – 8 volts is applied to the select gates SG_0 to SG_{15} ; 0 volts is applied to the erase gates EG_0 – EG_{15} and to select gate SG_{16} ; bit line BL_n is maintained at 0 volts; and the common source CS is biased at 0 volts. With these bias conditions, most of the applied voltage
15 appears across the gate oxide beneath the floating gate, which results in Fowler-Nordheim tunneling, with electrons migrating from the channel region to the floating gate. At the end of the program operation, the floating gate is negatively charged, and the threshold voltage of the memory cell, which preferably is in the range of 1 – 3 volts, becomes higher. Therefore, the memory
20 cell is turned off when the control gate is biased at 0 volt during a read operation. Following a program operation, the memory cell goes into a non-conductive state (logic "0").

In the unselected memory cells $C_{1(n-1)}$ and $C_{1(n+1)}$ that share the same control
25 gate CG_1 with the selected cell C_{1n} , the bit lines (BL_{n-1} and BL_{n+1}) are biased at 5 – 8 volts, and the control gate is biased at 15 – 18 volts. This results in negligible Fowler Nordheim tunneling in those cells, and the floating gate charges remain unchanged. In the other unselected memory cells C_{0n} and C_{2n} , the bit line BL_n is maintained at 0 volts, and 6 – 9 volts is applied to the control
30 gates (CG_0 and CG_2). This also minimizes Fowler-Nordheim tunneling, and the charges on the floating gates in those cells do not change either.

Another set of bias conditions for the program mode is illustrated in Figure 5B. In this example, the selected memory cell C_{1n} is biased with 10 – 13 volts on control
35 gate CG_1 , 0 – 3 volts is applied to select gates SG_0 to SG_{15} ; 0 or -5 volts is applied to erase gates EG_0 – EG_{15} ; -5 volts is applied to select gate SG_{16} , bit line BL_n and P-well 52; and the common source CS is biased at 0 volts. With the cell

biased in this manner, most of the applied voltage appears across the gate oxide beneath the floating gate. That results in Fowler-Nordheim tunneling, with electrons migrating from the channel region to the floating gate.

5 Figure 5C illustrates a set of bias conditions for the program mode with hot electron injection. The bias conditions are for selected memory cells on control gates with odd index numbers, e.g. CG_1, CG_3, CG_5 . For selected cell C_{1n} in Figure 5C, 10 – 12 volts is applied to control gate CG_1 ; 4 – 8 volts is applied to select gates SG_0 and $SG_2 - SG_{16}$; 0 volts is applied to erase gates $EG_0 - EG_{15}$;
10 4 – 8 volts is applied to selected bit line BL_n ; 7 – 9 volts is applied to the control gates of other memory cells in the same bit line direction as the selected cell (e.g. C_{0n}, C_{2n} , and C_{31n}); and the unselected bit lines (e.g. BL_{n-1} and BL_{n+1}) and common source CS are maintained at 0 volts. The voltage applied to the select gate (SG_1 in this example) just adjacent to the selected cell (C_{1n} in this example)
15 can be biased at the range of 1 – 2 volts. With these bias conditions, the cells and the select transistors are turned on.

Most of the voltage between the common source CS and the bit line BL_n appears across the mid-channel region between select gate SG_1 and the floating gate of the selected cell C_{1n} , resulting in a high lateral electric field in that region. In
20 addition, since the floating gate is coupled to a high voltage from bit line BL_n and control gate CG_1 , a strong vertical electric field is established near the split point of the select gate and the floating gate. When electrons flow from the common source to bit line during program operation, some of the channel electrons are
25 accelerated by lateral electric field, and some of the hot electrons are “hot” enough to exceed the energy barrier height between the channel and oxide (about 3.1 eV), and they will be injected into and collected on the floating gate due to the vertical field in the floating gate oxide. The injection point is near the split point of select gate and floating gate.

30 At the end of the program operation, the floating gate is negatively charged, and the threshold voltage of the memory cell, which is preferably on the range of 1 – 3 volts, becomes higher. Thus, the memory cell is turned off when the control gate is biased at 0 volts during a read operation. Following a program operation,
35 the memory cell goes into a non-conductive state (logic “0”)

In the unselected memory cells $C_{1(n-1)}$ and $C_{1(n+1)}$ which share the same control gate with the selected cell C_{1n} , the bit lines (BL_{n-1} and BL_{n+1}) are biased at 0 volts; the select gate SG_1 is at 1 – 2 volts; and the control gate CG_1 is at 10 – 12 volts.

5 The lateral voltage drop between the bit line and the common source is 0 volts, and there is no mid-channel hot carrier injection in cells $C_{1(n-1)}$ and $C_{1(n+1)}$. There is no hot electron injection in unselected memory cells such as C_{0n} , C_{2n} in the selected bit line because electrons flow from neighboring erase gate channels (under EG_0 and EG_1) to the cell channels. Cell C_{31n} is biased with 4 – 8 volts on both bit line BL_n and select gate SG_{16} , and 7 – 9 volts is applied to control gate
10 CG_{31} , which minimizes the mid-channel hot carrier injection, and the floating gate charges are unchanged.

In the read mode, the control gate CG_1 of the selected memory cell C_{1n} and the common source CS are biased to 0 volts; 1 – 3 volts is applied to bit line BL_n ; and
15 V_{cc} and 0 volt are applied to the select gates ($SG_0 - SG_{16}$) and erase gates ($EG_0 - EG_{15}$), respectively. The unselected memory cells in the bit line direction, e.g. C_{0n} and C_{2n} , are turned on by applying 5 – 8 volts to their control gates. When the memory cell is erased, the read operation shows a conductive state because the channel of selected cell is turned on. This is also the case in the other cells
20 and the select transistors in the same bit line direction. Thus, a logic “1” is returned by the sense amplifier. When the memory cell is programmed, the read shows a non-conductive state because the channel of the selected cell is turned off, and hence the sense amplifier returns a logic “0”. In the unselected memory cells $C_{1(n-1)}$ and $C_{1(n+1)}$, both the bit lines (BL_{n-1} and BL_{n+1}) and common source
25 CS are biased at 0 volts, and there is no current flow between the bit line and the common source nodes.

Figure 5D illustrates another set of bias conditions for the program mode utilizing hot electron injection. The bias conditions are for selected memory cell on
30 control gates with even index numbers, e.g. CG_0 , CG_2 , CG_4 . The main difference between the bias conditions of this figure and Figure 5C is that the bit line voltage and common source voltage are swapped in the program mode. For selected cell C_{2n} in Figure 5D, 10 – 12 volts is applied to control gate CG_2 ; 4 – 8 volts is applied to select gates SG_0 and $SG_2 - SG_{16}$; 0 volts is applied to erase
35 gates $EG_0 - EG_{15}$ and selected bit line BL_n ; 4 – 8 volts is applied to common source CS; 7 – 9 volts is applied to the control gates of other memory cells in the same bit line direction as the selected cell (e.g. C_{0n} , C_{1n} , and C_{31n}); and the

unselected bit lines (e.g. BL_{n-1} and BL_{n+1}) are biased at 3 volts. With these conditions, the cells and the select transistors are turned on, and the voltage applied to the select gate adjacent to the selected cell (SG_1 in this example) can be biased in the range of 1 – 2 volts.

5

Most of the voltage between the common source CS and the bit line BL_n appears across the mid-channel region between select gate SG_1 and the floating gate of the selected cell C_{2n} , resulting in a high lateral electric field in that region. In addition, since the floating gate is coupled to a high voltage from bit line BL_n and control gate CG_2 , a strong vertical electric field is established near the split point of select gate and floating gate. When electrons flow from the bit line to common source during programming, some of the channel electrons are accelerated by lateral electric field, and some of the hot electrons are “hot” enough to exceed the energy barrier height between the channel and oxide (about 3.1 eV), and they will be injected into and collected on floating gate by the vertical field in floating-gate oxide. The injection point is near the split point of select gate and floating gate.

10
15

At the end of the program operation, the floating gate is negatively charged, and the threshold voltage of the memory cell, which is preferably in the range of 1 – 3 volts, becomes higher. Thus, the memory cell is turned off when the control gate is biased at 0 volts during a read operation. Following a program operation, the memory cell goes into a non-conductive state (logic “0”).

20

The bit lines (BL_{n-1} and BL_{n+1}) for the unselected memory cells $C_{2(n-1)}$ and $C_{2(n+1)}$ which share the same control gate with the selected cell C_{2n} are biased at 3 volts, the select gate SG_1 is at 1 – 2 volts, and the control gate CG_2 is at 10 – 12 volts. Thus, select transistors $S_{1(n-1)}$ and $S_{1(n+1)}$ are turned off, and there is no mid-channel hot carrier injection in cells $C_{2(n-1)}$ and $C_{2(n+1)}$. In unselected memory cells such as C_{0n} , C_{1n} and C_{31n} in the selected bit line, there is no hot carrier injection. In cells C_{1n} and C_{31n} , electrons flow from neighboring erase gate channels (under EG_0 and EG_{15}) to the cell channels, and there is no mid-channel hot electron injection. Cell C_{0n} is biased with 4 – 8 volts on both common source gate CS and select gate SG_0 , and 7 – 9 volts is applied to control gate CG_0 , which minimizes the mid-channel hot carrier injection, and the floating gate charges are unchanged.

25

30

35

In the read mode, the bias conditions in Figure 5D are the same as in Figure 5C. The control gate of the selected memory cell C_{2n} and the source are maintained at 0 volts; 1 – 3 volts is applied to the bit line; and V_{cc} and 0 volts are applied to the select gates ($SG_0 - SG_{16}$) and erase gates ($EG_0 - EG_{15}$), respectively. The
5 unselected memory cells in the bit line direction, e.g. C_{0n} and C_{1n} , are turned on by applying 5 – 8 volts to their control gates. When the memory cell is erased, the read operation shows a conductive state because the channel of the selected cell is turned on, and that is also the case in the other cells and the select transistors in the same bit line direction. Thus, a logic “1” is returned by
10 the sense amplifier. When the memory cell is programmed, the read shows a non-conductive state because the channel of the selected cell is turned off, and hence the sense amplifier returns a logic “0”. In the unselected memory cells $C_{2(n-1)}$ and $C_{2(n+1)}$, both the bit line and common source nodes are biased at 0 volts, and there is no current flow between the bit line and the common source
15 nodes.

The embodiment of Figures 6 – 7 is generally similar to the embodiment of Figures 2 – 3, except the floating gates 37 are substantially thicker and do not have relatively sharp rounded edges in this embodiment. Control gates 38 cross
20 over floating gates 37 and the isolation regions 56 between them. Erase gates 43 and select gates 44, 44a, 44b extend in a direction perpendicular to the rows and parallel to the control gates. Bit lines 57 are perpendicular to the erase, select and control gates, and cross over the bit line contact 46, erase gates, select gates, and control gates 38 in each row of the array. The erase path
25 extends from the floating gate through tunnel oxide 40 to the channel region below.

A preferred process of fabricating the embodiment of Figures 6 - 7 is illustrated in Figures 8A - 8E. In this process, oxide layer 40 is thermally grown to a thickness
30 of about 60Å to 120Å on a monocrystalline silicon substrate which, in the embodiment illustrated, is in the form of a P-type substrate 41 in which a P-type well 52 is formed. Alternatively, if desired, an N-type well can be formed in the P-type substrate, in which case the P-type well will be formed in the N-type well.

35 A conductive layer 62 of polysilicon or amorphous silicon (poly-1) is deposited on the thermal oxide to a thickness on the order of 300Å to 1500Å, and an inter-poly dielectric layer 42 is formed on the silicon. The silicon is preferably doped with

phosphorus, arsenic or boron to a level on the order of 10^{17} to 10^{20} per cm^3 . The doping can be done *in-situ* during deposition of the silicon or by ion implantation either directly into the silicon or through the dielectric 42 above it. The inter-poly dielectric can be either a pure oxide or a combination of oxide, nitride and oxide (ONO), and in the embodiment illustrated, it consists of a lower oxide layer having a thickness on the order of 30\AA - 100\AA , a central nitride layer having a thickness on the order of 60\AA - 200\AA , and an upper oxide layer having a thickness on the order of 30\AA - 100\AA .

A second layer 63 of polysilicon (poly-2) is deposited on dielectric film 42. This layer has a thickness on the order of 1500\AA - 3500\AA , and is doped with phosphorus, arsenic or boron to a level on the order of 10^{20} to 10^{21} per cm^3 . A CVD oxide or nitride layer 66 having a thickness on the order of 300\AA - 1000\AA is deposited on the poly-2 layer, and is used as a mask to prevent the poly-2 material from etching away during subsequent dry etching steps.

A photolithographic mask 67 is formed over layer 66 to define the control gates, and the unmasked portions of that layer and poly-2 layer 63 are etched away anisotropically, leaving only the portions of the poly-2 which form the control gates 38. The exposed portions of the inter-poly dielectric 42 and the underlying portions of the poly-1 layer 62 are then etched away anisotropically to form the floating gates 37, as illustrated in Figure 8B. Thereafter, diffusion regions 49 are formed in the substrate between the stack gates by ion implantation using with dopants such as P^{31} or As^{75} .

Following ion implantation, a dielectric 47 is formed on the sidewalls of control and floating gates, and a conductive (poly-3) layer 59 is deposited over the entire wafer, as shown in Figure 8C. The dielectric can be either a pure oxide or a combination of oxide, nitride and oxide (ONO), and in the embodiment illustrated, it consists of a lower oxide layer having a thickness on the order of 30\AA - 100\AA , a central nitride layer having a thickness on the order of 60\AA - 300\AA , and an upper oxide layer having a thickness on the order of 30\AA - 100\AA . The poly-3 layer is typically doped polysilicon or polycide, and is deposited to a thickness on the order of 1500\AA - 3000\AA .

The poly-3 layer is then etched anisotropically to form erase gates 43 and select gates 44, 44a, 44b, as illustrated in Figure 8D. Being formed in this manner, the

erase and select gates are self-aligned and parallel to the control gates. N-type dopants such as P^{31} or As^{75} are implanted into P-well 52 to form the bit line diffusion 50 and common source diffusion 51.

5 Thereafter, a glass material 60 such as phosphosilicate glass (PSG) or borophosphosilicate glass (BPSG) is deposited across the entire wafer, then etched to form openings for bit line contacts 46, as shown in Figure 8E. Finally, a metal layer is deposited over the glass and patterned to form bit lines 57 and bit line contacts 46.

10

Operation of the embodiment of Figures 6 - 7 is generally similar to that of the embodiment of Figures 2 - 3, and exemplary bias voltages for erase (ERS), program (PGM) and read (RD) operations are shown next to the terminals of the array in Figures 9A - 9B. In this example, memory cell C_{1n} is once again selected. This cell is located at the intersection of control gate CG_1 and bit line BL_n , and is encircled on the drawing for ease of location. All of the other memory cells in the array are unselected.

15

During an erase operation, electrons are forced to tunnel from the floating gate to the channel region beneath it, leaving positive ions in the floating gate. When the electric field across the tunnel oxide is more than 10 mV/cm, Fowler-Nordheim tunneling becomes significant, and electrons with sufficient energy can tunnel from the floating gate to the channel region.

20

25 With the control gate, erase gate and select gate surrounding the floating gate or cathode electrode, high-voltage coupling from the control gate, erase gate and select gate to the floating gate is once again substantially enhanced, and the voltage required for Fowler-Nordheim tunneling is reduced significantly. The enhanced coupling also makes it possible to use a thicker tunnel oxide while still maintaining sufficient electron tunneling.

30

Erasing can be done using either of two bias conditions. In the first erase mode (ERS1), the control gate is biased at a level on the order of -11 to -18 volts, the select gates SG_0 to SG_{16} and erase gates EG_0 - EG_{15} are biased at -6 to -13 volts, and the bit line, common source and P-well are biased at 0 volts. In the second erase mode (ERS2), the control, erase and select gates are biased at 0

35

volts, the bit line and common source are floating, and the P-well is biased at 10 to 13 volts.

5 With these bias conditions, most of the voltage applied between the control gate and the select gates appears across the tunnel oxide under the floating gate. That triggers Fowler-Nordheim tunneling, with electrons tunneling from the floating gate to the underneath channel region. As the floating gate becomes more positively charged, the threshold voltage of the memory cell, which is preferably on the order of -2 to -5 volts in this embodiment, becomes lower. That
10 results in an inversion layer in the channel under the floating gate when the control gate is biased at 0 volts. Therefore, the memory cell goes into the conductive state (logic "1") after the erase operation.

15 In the unselected memory cells, the control gates, erase gates and select gates are biased at 0 volts, so there is no Fowler-Nordheim tunneling in them during the erase operation.

During a programming operation for selected memory cells on control gates with odd index numbers, e.g. CG_1 , CG_3 , CG_5 , the control gate of the selected
20 memory cell C_{1n} is biased to a level of 9 - 11 volts, 4 - 8 volts is applied to select gates SG_0 and $SG_2 - SG_{16}$, 0 volts is applied to erase gates $EG_0 - EG_{15}$, 7 - 11 volts is applied to the control gates of the other memory cells in the same bit line direction as the selected cell (e.g. C_{0n} and C_{2n}), the common source and P-well are held at 0 volts, and 4 - 8 volts is applied to the bit line. The cells and the
25 select transistors are turned on by applying 7 - 11 volts to the control gates and 4 - 8 volts to the select gates. The voltage applied to the select gate just before the selected cell (SG_1 and C_{1n} in this example) can be on the low side, preferably on the order of 1 - 2 volts.

30 With these bias conditions, most of the voltage between the common source and the bit line appears across the mid-channel region between select gate SG_1 and the floating gate of the selected cell C_{1n} , resulting in a high electric field in that region. In addition, since the floating gate is coupled to a high voltage from the common source node (i.e., control gate CG_1 and select gate SG_2), a strong
35 vertical electric field is established across the oxide between the mid-channel region and the floating gate. When electrons flow from the bit line to the common source during the program operation, they are accelerated by the electric field

across the mid-channel region, and some of them become heated. Some of the hot electrons get accelerated by the vertical field, which causes them to overcome the energy barrier of the oxide (about 3.1 eV) and inject into the floating gate.

5

At the end of the program operation, the floating gate is negatively charged, and the threshold voltage of the memory cell, which preferably is on the order of 2 - 4 volts, becomes higher. Thus, the memory cell is turned off when the control gate is biased at 0 volts during a read operation. Following a program operation, the

10

The bit line for the unselected memory cells $C_{2(n-1)}$ and $C_{2(n+1)}$ which share the same control gate with the selected cell C_{1n} is biased at 3 volts, the select gate SG_1 is at 1 - 2 volts, and the control gate is at 9 - 11 volts. Thus, select transistors $S_{1(n-1)}$ and $S_{1(n+1)}$ are turned off, and there is no mid-channel hot carrier injection in cells $C_{1(n-1)}$ and $C_{1(n+1)}$. In unselected memory cells such as C_{0n} , C_{2n} and C_{31n} in the selected bit line, there is no hot carrier injection. In cells C_{0n} and C_{2n} , electrons flow from neighboring erase gate channels (under EG_0 and EG_1) to the cell channels, and there is no mid-channel hot electron injection.

15

Cell C_{31n} is biased with 4 - 8 volts to both the bit line BL_n and select gate SG_{16} , and 7 - 9 volts to the control gates CG_{31} , which minimizes the mid-channel hot carrier injection, and the floating gate charges are unchanged.

20

In the read mode, the control gate of the selected memory cell C_{1n} is biased at 0 - 1.5 volts, the common source is biased to 0 volts, 1 - 3 volts is applied to the bit line, V_{cc} is applied to the select gates $SG_0 - SG_{16}$, and 0 volts is applied to the erase gates $EG_0 - EG_{15}$. The unselected memory cells in the bit line direction, e.g. C_{0n} and C_{2n} , are turned on by applying 5 - 9 volts to their control gates. When the memory cell is erased, the read shows a conductive state because the channel of selected cell is turned on, and the other cells and the select transistors in the same bit line direction also turned on. Thus, a logic "1" is returned by the sense amplifier. When the memory cell is programmed, the read shows a non-conductive state because the channel of the selected cell is turned off, and hence the sense amplifier returns logic "0". In the unselected memory cells $C_{1(n-1)}$ and $C_{1(n+1)}$, both the bit line and common source nodes are biased at 0 volts, and there is no current flow between the bit line and the common source nodes.

25

30

35

Figure 9B illustrates the bias conditions for selected memory cell on control gates with even index numbers, e.g. CG_0 , CG_2 , CG_4 . The main difference in the bias conditions shown in Figures 9A and 9B is that the bit line voltage and common source voltage are swapped in program mode. For selected cell C_{2n} in Figure 9B, 9 – 11 volts is applied to the control gate CG_2 ; 4 – 8 volts is applied to select gates SG_0 and $SG_2 - SG_{16}$; 0 volt is applied to erase gates $EG_0 - EG_{15}$ and selected bit line BL_n ; 4 – 8 volts is applied to common source CS ; 7 – 11 volts is applied to the control gates of other memory cells in the same bit line direction as the selected cell (e.g. C_{0n} , C_{1n} , and C_{31n}); and the unselected bit lines (e.g. BL_{n-1} and BL_{n+1}) are biased at 3 volts. The cells and the select transistors are turned on these voltages. The voltage applied to the select gate (SG_1 in this example) adjacent to the selected cell (C_{2n} in this example) can be biased to about 1 – 2 volts.

With these bias conditions, most of the voltage between the common source CS and the bit line BL_n appears across the mid-channel region between select gate SG_1 and the floating gate of the selected cell C_{2n} , resulting in a high lateral electric field in that region. In addition, since the floating gate is coupled to a high voltage from bit line BL_n and control gate CG_2 , a strong vertical electric field is established near the split point of select gate and floating gate. When electrons flow from the bit line to common source during programming, some of the channel electrons are accelerated by the lateral electric field, and some of the hot electrons are "hot" enough to surmount energy barrier height for electron between channel and oxide (about 3.1 eV), and they will be injected into and collected on floating gate because of the vertical field in floating-gate oxide. The injection point is near the split point of select gate and floating gate.

At the end of the program operation, the floating gate is negatively charged, and the threshold voltage of the memory cell, which is preferably in the range of 1 – 3 volts, becomes higher. Thus, the memory cell is turned off when the control gate is biased at 0 volts during a read operation. Following a program operation, the memory cell goes into a non-conductive state (logic "0")

For the unselected memory cells $C_{2(n-1)}$ and $C_{2(n+1)}$ which share the same control gate with the selected cell C_{2n} , bit lines (BL_{n-1} and BL_{n+1}) are biased at 3 volts; the select gate SG_1 is at 1 – 2 volts; and the control gate CG_2 is at 9 – 11 volts.

Thus, select transistors $S_{1(n-1)}$ and $S_{1(n+1)}$ are turned off, and there is no mid-channel hot carrier injection in cells $C_{2(n-1)}$ and $C_{2(n+1)}$. In unselected memory cells such as C_{0n} , C_{1n} , and C_{31n} in the selected bit line, there is no hot carrier injection. Electrons flow from the erase gate channels adjacent to cells C_{1n} and C_{31n} (under EG_0 and EG_{15}) to cell channels; and thus there is no mid-channel hot electron injection. Cell C_{0n} is biased with 4 – 8 volts on both the common source CS and the select gate SG_0 , and 7 – 11 volts on the control gates CG_0 , which minimizes the mid-channel hot carrier injection, and the floating gate charges are unchanged.

10

In the read mode, the bias conditions shown in Figures 9A and 9B are the same. The control gate of the selected memory cell C_{2n} and the source are biased to 0 – 1.5 volts; 1 – 3 volts is applied to the bit line; and V_{cc} and 0 volts are applied to the select gates ($SG_0 - SG_{16}$) and erase gates ($EG_0 - EG_{15}$), respectively. The unselected memory cells in the bit line direction, e.g. C_{0n} and C_{1n} , are turned on by applying 5 – 9 volts to their control gates. When the memory cell is erased, the read operation shows a conductive state because the channel of selected cell is turned on. This is also the case for the other cells and the select transistors in the same bit line direction. Thus, a logic "1" is returned by the sense amplifier. When the memory cell is programmed, the read shows a non-conductive state because the channel of the selected cell is turned off, and hence the sense amplifier returns a logic "0". In the unselected memory cells $C_{2(n-1)}$ and $C_{2(n+1)}$, both the bit line and common source nodes are biased at 0 volts, and there is no current flow between the bit line and the common source nodes.

25

The invention has a number of important features and advantages. It provides a self-aligned split-gate NAND flash memory cell array which has significantly smaller cell size and greater cell density than memory structures heretofore provided. The control and floating gates in each cell are stacked and self-aligned with each other, and the erase gates and select gates are split from but self-aligned with the stacked gates. Resistance of the channel region between bit line diffusion and common source region is reduced significantly by diffusions beneath the erase gates, which permits the length of the structure and the number of cells in each row to be substantially greater than in devices which do not have such diffusions. In addition, the control gates, the select gates and the erase gates surround the floating gates in a manner which provides a

30

35

relatively large inter-gate capacitance for high-voltage coupling during program and erase operations.

5 It is apparent from the foregoing that a new and improved self-aligned split-gate NAND flash memory and process of fabrication have been provided. While only certain presently preferred embodiment has been described in detail, as will be apparent to those familiar with the art, certain changes and modifications can be made without departing from the scope of the invention as defined by the following claims.